

Tema 3: Alineamientos pareados y búsqueda de homólogos en bases de datos mediante BLAST

Introducción a la filoinformática – pan-genómica y filogenómica microbiana, TIB2025, 4-8 de agosto, 2025 CCG-UNAM, Cuernavaca, Mor. México <https://github.com/vinuesa/TIB-filoinfo> CCG-UNAM

Introducción a la Filoinformática: Pan-genómica y Filogenómica microbiana– NNB & CCG-UNAM, 4-8 de agosto 2025

Pablo Vinuesa ([vinuesa\[at\]ccg.unam.mx](mailto:vinuesa[at]ccg.unam.mx); [@pvinmex](https://twitter.com/pvinmex))
Centro de Ciencias Genómicas, CCG-UNAM, Cuernavaca, México
<http://www.ccg.unam.mx/~vinuesa/>

Todo el material del curso (presentaciones, tutoriales y datos) lo encontrarás en:
<https://github.com/vinuesa/TIB-filoinfo>

• **Tema 3: alineamientos pareados y búsqueda de homólogos en bases de datos mediante BLAST**

- evolución de secuencias y **clasificación de mutaciones**
- **indeles y gaps**
- **alineamientos globales** (Needleman-Wunsch) vs. **locales** (Smith-Waterman);
- **matrices de costo de sustitución, penalización de gaps** y cuantificación de la similitud;
- evaluación estadística de la similitud entre pares de secuencias;
- escrutinio de bases de datos mediante **BLAST**; Búsquedas a nivel de **DNA vs. AA**;
- la **familia BLAST** e interpretación de resultados de **búsqueda de secuencias homólogas**
- prácticas: uso de **NCBI BLAST desde la línea de comandos**



Protocolo básico para un análisis filogenético de secuencias moleculares

Tema 2:
alineamientos pareados, búsquedas de homólogos en bases de datos

Colección de secuencias homólogas
• **BLAST, diamond**

Alineamiento múltiple de secuencias
• **clustalo, mafft, muscle ...**

Análisis evolutivo del alineamiento y selección del modelo de sustitución más ajustado
• **tests de saturación, modeltest, ...**

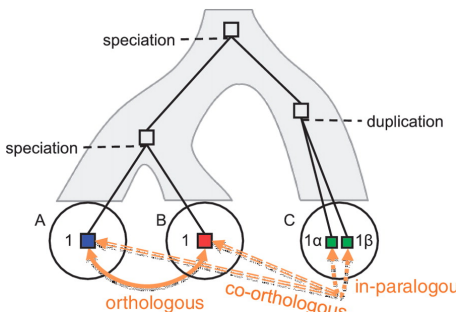
Estima filogenética
• **NJ, ME, MP, ML, Bayes ...**

Pruebas de confiabilidad de la topología inferida
• **proporciones de bootstrap probabilidad posterior ...**

Interpretación evolutiva y aplicación de las filogenias

Basic concepts in bacterial genome evolution – homology relationships between gene families

Orthology, co-orthology and paralogy relationships in the evolution of four genes that arose from a single common ancestor.



Gene trees are not the same as species trees!!!

Kristensen D M et al. Brief Bioinform 2011;12:379-391

Orthologs: essentially instances of 'the same gene' in different species. Tend to retain function across large phylogenetic distances. Key markers for phylogenetics and genome annotation.

Paralogs: Tend to diverge over time to perform different functions via subfunctionalization or neofunctionalization routes.

Homología entre secuencias de DNA y proteína: tipos de mutaciones en secs. codificadoras de proteínas

secuencia ancestral
pos. codón 123
codones **ATG TGT TTT GAT GCA**
AA M C F D A

secuencias derivadas (evolucionadas)

especie A
ATG TAT TTT CAT GCA
M T F H A
no-sinónima

especie B
ATG --- TTC GAC GCA
M F D A
sinónimas y delección en marco

especie C
ATG TGT TT- G ATG CAX
M C L M X
delección fuera de marco

- Todas las mutaciones en 2^{as} posiciones resultan en sustituciones no sinónimas
- 96% de mutaciones en 1^{as} posiciones resultan en sustituciones no sinónimas
- Casi todas las sustituciones sinónimas ocurren en las 3^{as} posiciones
- las delecciones o inserciones en secs. codificadoras de aa suceden generalmente en múltiplos de tres nt; de no ser así se generan cambios de marco de lectura corriente abajo de la mutación, con frecuencia generando un pseudogen no funcional

Homología entre secuencias de DNA y proteína:
alineamiento y tipos de mutaciones

secuencia ancestral

pos. codón 123
codones ATG TGT TTT GAT GCA
AA M C F D A

alineamiento de
sitios homólogos
para tres secs.

especie A ATG TAT TTT CAT GCA
especie B ATG --- TTC GAC GCA
especie C ATG TGT TT- GAT GCA

cambio de marco de
lectura !!! posible
pseudogen.

Transiciones (ti) purina - purina

A

G

β_{A-C}

C

T

α_{C-G}

Transiciones (ti) pirimidina - pirimidina

Transversiones (tv)
pur. <-> pyr.

• existen 4 tipos de ti y 8 de tv

• las tasas de sustitución de ti (↔) son
generalmente mucho más altas que las
de tv (↕)

Computational methods to identify homologs

Methods to identify orthologs

Phylogenetic
Tree-based
reconciliation
methods

• Orthostrapper
• RIO
• COCO-CL
• TreeFam
• PHOG
• MetaPhOrs

Blast-based
heuristic
best-match
methods

• BDBHs
• COGs
• OrthoMCL

• Most direct approach
but known to be
subjected to increasing
artifacts as evol. dist.
increases due to:
- alignment problems
- long-branch attraction
- heterotachy
- tree searches stuck
in local sub-optimal
maxima

• More adequate
for bacteria, as
they don't follow
a clear tree-like
evolutionary path
(HGT).

• The only practical
strategy for large
datasets, therefore
most common
method in
comparative
genomics

Computational methods to identify homologs

Grouping of genes in different species that are each others' BLAST
best bi-directional hits (BBHs) into sets of orthologs and co-orthologs.

Linking pairs of BBHs from
multiple genomes has a property
of self-verification, as their consis-
tency would be very unlikely due to
chance, especially between phylo-
genetically distant lineages.

Methods for clustering of pairwise
BBHs vary, but the most widely
used approach involves a single-
linkage clustering procedure, where any
two clusters sharing a common BBH are
merged until convergence.

Problems with: differential gene loss,
domain recombination/gain/loss

(a)

(b)

(c)

Species

Genes / Domains

SymBet

Sets of in-paralogs

Kristensen D M et al. Brief Bioinform 2011;12:379-391

Alineamientos pareados y búsqueda de homólogos en bases de datos

Los alineamientos pareados son la base de los métodos de búsqueda de
secuencias homólogas en bases de datos

• Si dos proteínas o genes se parecen mucho a lo largo de toda su longitud, asumimos
que se trata de proteínas o genes homólogos, es decir, descendientes de un mismo
ancestro común (cenancestro).

• Por ello, una de las técnicas más utilizadas para detectar potenciales homólogos en
bases de datos de secuencias se basa en la **cuantificación de la similitud entre pares
de secuencias**, y la determinación de la **significancia estadística** de dicho parecido.
Estas magnitudes son las que reportan los estadísticos de **BLAST**.

>gi171548896|ref|ZP_00669120.1| Translation elongation factor G:Small GTP-binding protein domain
[Nitrosomonas eutropha C71]
gi171486077|gb|EAO18626.1| Translation elongation factor G:Small GTP-binding protein domain
[Nitrosomonas eutropha C71]
Length=696

Score = 828 bits (2140), Expect = 0.0
Identities = 434/697 (62%), Positives = 541/697 (77%), Gaps = 9/697 (1%)

Query 1 MTREFSLEKTRNIGIMAHIDAGKTTTTERVLYTTGRIKIGETHEGASQMDWMAQEQERG 60
M++ LE+ RNIGIMAHIDAGKTTT+ER+L+YTG HK+GE H+GA+ MDWM QEQERG 60
Sbjct 1 MSKRNP LERYRNIGIMAHIDAGKTTTTERILFTYTGVS HKLGEVHDGAATMDWMEQEQERG 60

Query 61 XXXXXXXXXXXXXN-----DHRINIIDTPGHVDFTVEVERSLRVLDGAVAVLDAQSGVE 113
TTTSAATT W +HRIN+IDTPGHVDFT+EVERSLRVLDGA V + GV+ 113
Sbjct 61 TTTSAATTCTFVKGAGNTPFHRINVIDTPGHVDFTIEVERSLRVLDGACTVFCVSGGVQ 120 (... truncado)

© Pablo Vinuesa 2025. @pvinmex; vinuesa[at]ccg[dot]unam[dot]mx;
<http://www.ccg.unam.mx/~vinuesa/>

Licencia Creative Commons 4.0, no comercial
Attribution-NonCommercial 4.0 International (CC BY-NC 4.0)
<https://creativecommons.org/licenses/by-nc/4.0/>

Tema 3: Alineamientos pareados y búsqueda de homólogos en bases de datos mediante BLAST

Programación dinámica y la generación de alineamientos pareados (globales y locales)

- Pares de secuencias pueden ser comparadas usando **alineamientos globales y locales**, dependiendo del objetivo de la comparación.

Un **alineamiento global** fuerza el alineamiento de ambas secuencias a lo largo de toda su longitud. Usamos aln. globales cuando estamos seguros de que la homología se extiende a lo largo de todas las secuencias a comparar. Este es el tipo de alineamientos que generan programas de **alineamiento múltiple** tales como clustal, mafft o muscle.

(a)

P00001	1	MGDVEKGKKIFIMKCSQCHTVEKGGKHKGTGPNLHGLFGRKGTQAPGYSYTAANKNK---GI	58
		D KG+ +F QC T + K+ GP L G+ GRK G A G++Y+ N N G+	
P00090	1	Q-DAAKGEAVF-----KQCMTCHRADKNMVGPA LGGVVGKAGTAAGFTYSPLNHNHSGEAGL	56
P00001	59	IWGEDTLMHYLENPKKYIP-----GTMKIFVGIIKKKEERADLLIAYLKKATNE	105
		+W ++ ++ YL +P Y+ T M F + ++R D+ AYL AT +	
P00090	57	VWTQENIIAYLPDPNAYLKKFLTDKGQADKATGSTKMTF-KLANQQRKDVAAYL--ATLK	114

Alineamiento global óptimo del citocromo C humano (105 **resíduos**, SWISS-PROT acc. P00001) y citocromo C2 de Rhodopseudomonas palustris (114 **resíduos**, SWISS-PROT acc. P00090).

La **matriz de puntuación o ponderación** ("scoring matrix") empleada fue **BLOSUM62**, con **costo de gaps afines** de $-(11 + k)$. La puntuación del alineamiento global es de 131, usando el algoritmo de **Needleman-Wunsch**.

Introducción a la filoinformática – pan-genómica y filogenómica microbiana, TIB2025, 4-8 de agosto, 2025 CCG-UNAM, Cuernavaca, Mor. México <https://github.com/vinuesa/TIB-filoinfo> CCG-UNAM

Programación dinámica y la generación de alineamientos pareados (globales y locales)

Un **alineamiento local** sólo busca los segmentos con la puntuación más alta. Se usa, por ejemplo, en el **escrutinio de bases de datos** de secuencias debido a que la homología entre pares de secuencias frecuentemente existe sólo a nivel de ciertos dominios, pero no a lo largo de toda la secuencia (**estructura modular de genes y proteínas**). **BLAST y diamond** buscan alineamientos locales con alta puntuación (**HSPs** ó high-scoring pairs)

(b)

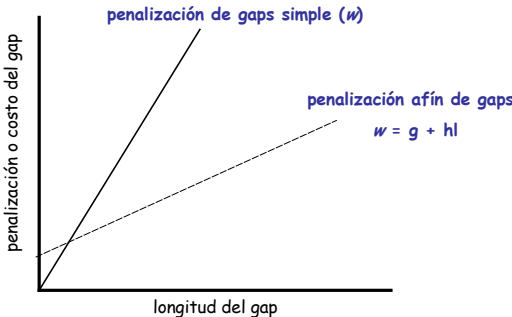
P13569	1221	EGGNAILNIFSISIPGQVVLGRTSGGKSTLLSAFLRL-----NTEGEIQIDGVS	1273
		+ ++ +S ++ G+ + L+G +SGKS +A L +L T GEI DG	
P33593	13	QAAQPLVHGVSLLTQGRVLA L VGGSGGKSLTCAATLGILPAGVQTAGEILADGKP	70
P13569	1274	WDSITL-----QQWRKAFGVIPQKVFIFSGTFRKNLDPYEQWSDQEIWKVADEV	1322
		L Q R AF + + + + + K AD+	
P33593	71	VSPCALRGIKIATIMQNPFRSAFNPL-----HTMHTARETCLALGKPADDA	116
P13569	1323	GLRSVIEQFP-GKLDVFLVDGGCVLSHGKQLMCLARSLVSKAKILLDEPSAHLDPV	1379
		L + IE VL +S G Q M +A +VL ++ ++ DEP+ LD V	
P33593	117	TLTAATEAVGLENAARVLKLYPFEMSGMLQRMMIAMVLCESPFPIADEPTDLDV	174

Alineamiento local óptimo del regulador de conductancia transmembranal de fibrosis cística de humano (1480 **resíduos**, SWISS-PROT acc. P13569) y la proteína transportadora de Ni dependiente de ATP de E. coli (253 **resíduos**, SWISS-PROT acc. P33593).

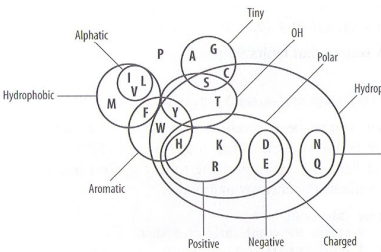
La **matriz de puntuación o ponderación** ("scoring matrix") empleada fue **BLOSUM62**, con **costo de gaps afines** de $-(11 + k)$. La puntuación del alineamiento local es de 89, usando el algoritmo de **Smith-Waterman**.

alineamientos pareados y factores de penalización afines para gaps

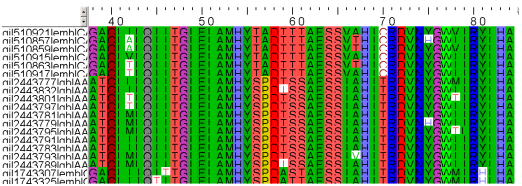
- Dado que **un sólo evento mutacional puede insertar o eliminar varios nucleótidos** de una secuencia, un **indel** largo no debe de ser penalizado proporcionalmente más que otro más corto ubicado en la misma región de un gen.
De ahí el uso de **factores de penalización afines para gaps** (affine gap penalties or costs (**w**), que cobran una penalidad relativamente alta por abrir un gap (**g**) y una penalidad más baja (**h**) por cada posición (**l**) sobre la que se extiende.
- La calidad de un alineamiento depende en gran medida de los **valores de apertura y extensión de gap** elegidos.



Cuantificación de la similitud entre pares de secuencias de AA



- Este diagrama muestra a los **aminoácidos** agrupados atendiendo a sus características químicas y físicas.
- Desde una perspectiva evolutiva **esperamos encontrar más sustituciones entre aas. similares que entre los menos relacionados**.
- Estos patrones pueden observarse en alineamientos múltiples como el mostrado abajo



Cuantificación de la similitud entre pares de secuencias de AA

• Las matrices empíricas de sustitución entre AAs no reflejan necesariamente las relaciones químicas entre ellos. Se trata de una **definición puramente estadística, basada en el análisis de frecuencias empíricas de sustituciones observadas en alineamientos de secs. con un grado de divergencia definido**

Cada **score** de la matriz **representa la tasa de sustitución esperada entre un par de AAs**. Por tanto, los **scores de los alineamientos pareados** evaluados con estas matrices **reflejan la similitud evolutiva existente entre las secuencias**.

Es importante notar que los **scores son evolutivamente simétricos** al no conocerse la dirección del cambio evolutivo (rev. temp.)

Table 2 - The log odds matrix for BLOSUM 62

A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	
4	0	-2	-1	-3	0	-2	-1	-4	-3	-2	-3	-1	-2	-3	-1	-1	-4	-4	-2	
C		9	-3	-4	-1	-3	-3	-1	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	
D			6	5	-2	0	0	1	1	0	0	0	0	0	0	0	0	0	0	
E				6	-2	4	-1	0	0	0	0	0	0	0	0	0	0	0	0	
F					6	3	-1	0	0	0	0	0	0	0	0	0	0	0	0	
G						6	-1	0	0	0	0	0	0	0	0	0	0	0	0	
H							6	-1	0	0	0	0	0	0	0	0	0	0	0	
I								4	-3	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	
K									4	-1	0	0	0	0	0	0	0	0	0	
L										6	-2	-1	-1	-1	-1	-1	-1	-1	-1	
M											6	-2	-1	-1	-1	-1	-1	-1	-1	
N												6	-2	-1	-1	-1	-1	-1	-1	
P													6	-2	-1	-1	-1	-1	-1	
Q														6	-2	-1	-1	-1	-1	
R															6	-2	-1	-1	-1	
S																6	-2	-1	-1	
T																	6	-2	-1	
V																		6	-2	
W																			6	
Y																				6

Matriz BLOSUM62

Cuantificación de la similitud entre pares de secuencias de AA

• **Matrices de sustitución de AAs log-odds (lod) scores (razón de probabilidades)**

$$s(a,b) = (c) \log \frac{p_{ab}}{f_a f_b}$$

$s(a,b)$ = **score crudo** del par a, b
Refleja la tendencia relativa de encontrarlos alineados

P_{ab} = verosimilitud de la hipótesis a evaluar; **frecuencia esperada o diana**, probabilidad con la que esperamos encontrar a y b apareados en un alineam. múltiple; es la que **observamos empíricamente**.

$f_a f_b$ = verosimilitud de la hipótesis nula; **frecuencia de fondo**, probabilidad con la que esperamos encontrar a y b en cualquier proteína. Refleja su abundancia o frecuencia

c = **Factor de escalamiento** usado para multiplicar los *lod scores* (números reales) antes de ser redondeados a números enteros, tal y como se observa en la matriz.
Los valores enteros redondeados resultantes se conocen como **"raw scores/puntajes crudos"**.

Table 2 - The log odds matrix for BLOSUM 62

A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	
4	0	-2	-1	-3	0	-2	-1	-4	-3	-2	-3	-1	-2	-3	-1	-1	-4	-4	-2	
C		9	-3	-4	-1	-3	-3	-1	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	
D			6	5	-2	0	0	1	1	0	0	0	0	0	0	0	0	0	0	
E				6	-2	4	-1	0	0	0	0	0	0	0	0	0	0	0	0	
F					6	3	-1	0	0	0	0	0	0	0	0	0	0	0	0	
G						6	-1	0	0	0	0	0	0	0	0	0	0	0	0	
H							6	-1	0	0	0	0	0	0	0	0	0	0	0	
I								4	-3	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	
K									4	-1	0	0	0	0	0	0	0	0	0	
L										6	-2	-1	-1	-1	-1	-1	-1	-1	-1	
M											6	-2	-1	-1	-1	-1	-1	-1	-1	
N												6	-2	-1	-1	-1	-1	-1	-1	
P													6	-2	-1	-1	-1	-1	-1	
Q														6	-2	-1	-1	-1	-1	
R															6	-2	-1	-1	-1	
S																6	-2	-1	-1	
T																	6	-2	-1	
V																		6	-2	
W																			6	
Y																				6

Matriz BLOSUM62

Estadísticos de Karlin-Altschul de similitud entre secuencias: frecuencias diana, lambda y entropía relativa

Los atributos más importantes de una matriz de sustitución son sus **frecuencias esperadas o diana** implícitas para cada par de aa en sus respectivos **scores crudos**. Estas frecuencias esperadas **representan el modelo evolutivo subyacente, resumido en la matriz**. Los scores que han sido re-escalados y redondeados (scores representados en la matriz) son los **scores crudos** s_{ab} . Para convertirlos a un **score normalizado** (log-odd score original) tenemos que multiplicarlos por λ , una constante específica para cada matriz. λ es aprox. igual al inverso del factor de escalamiento (c).

$$s(a,b) = \frac{1}{\lambda} \log \frac{p_{ab}}{f_a f_b} \quad p_{ab} = f_a f_b e^{\lambda s_{ab}} = \text{score normalizado}$$

por tanto, para despejar λ necesitamos $f_a f_b$ y encontrar el valor de λ para el que la suma de las frecuencias diana implícitas valga 1, ya que son frecuencias relativas.

$$\sum_{a=1}^n \sum_{b=1}^n p_{ab} = \sum_{a=1}^n \sum_{b=1}^n f_a f_b e^{\lambda s_{ab}} = 1$$

Una vez calculada λ , se usa para calcular el **valor de expectancia (E)** de cada **HSP (High Scoring Pair)** en el reporte de una búsqueda **BLAST**

Dado que las $f_a f_b$ de los residuos de algunas proteínas difieren mucho de las frecuencias de residuos empleadas para calcular las matrices PAM y BLOSUM, versiones recientes de BLASTP y PSI-BLAST incorporan una **"composition-based λ "** que es **"hit-específica"**

Estadísticos de Karlin-Altschul para alineamientos locales

Karlin, S., and Altschul, S. F. 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc Natl Acad Sci U S A **87**: 2264-268.
<https://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html>

Para evaluar si un alineamiento pareado (HSP) representa evidencia de homología es útil calcular con qué fuerza cabe esperar dicho alineamiento por simple azar. Esto es lo que calcula la ecuación de K-A.1

Esta ecuación indica que el **número de HSPs con score $\geq S$ esperados por azar (E)** durante una búsqueda de similitud en una base de datos de secuencias y es función directa de:

1. el tamaño del espacio de búsqueda (m, n),
2. el **score normalizado del HSP = score crudo (S) * λ** , donde λ se calcula a partir de la matriz de sustitución usada)
3. una constante de escalamiento de valor pequeño (k) para el espacio de búsqueda

E Describe el ruido de fondo, por azar, presente en el **hit** encontrado

m = número de símbolos en la secuencia problema
 n = número de símbolos en la base de datos
 $k \approx 0.1$ constante de ajuste para considerar HSPs altamente correlacionados
 λ = constante de escalamiento de la matriz ($\approx 1/c$)

Tema 3: Alineamientos pareados y búsqueda de homólogos en bases de datos mediante BLAST

Estadísticos de Karlin-Altschul para alineamientos locales
- ejemplo para un HSP de 10 nucleótidos, usando matriz +2/-1

$$E = k m n e^{-\lambda S}$$

Dadas la matriz de puntuación: **match = +1; mismatch = -1**
y el siguiente HSP (alinamiento local) de 10 residuos con 60% de identidad

Query: ACGTACGTAC

| | | | |

Subject: ATGTTCATGC

Calculamos el **score crudo** $S = \sum \text{substitution scores} = 6 \times 1 + 4 \times (-1) = 6 - 4 = 2$

Dados los siguientes valores:

$k = 0.1$; $m = 10$ (longitud de la secuencia problema o query); $n = 100$ (tamaño en residuos de la base de datos (número hipotético))

Sabiendo que $\lambda = 1.58$ para esta matriz (calculado de la matriz de ponderación), calculemos E .

$$E = 0.1 \times 10 \times 100 \times e^{-1.58 \times 2} \approx 100 \times e^{-3.16} \approx 100 \times 0.0424 = 4.24;$$

Donde $P = 1 - e^{-E} = 0.98$. Es por tanto un hit fortuito, que no representa evidencia de homología

Cómputo de estadísticos de Karlin-Altschul para alineamientos locales
- ejemplo para dos HSP de 10 y 300 nucleótidos, usando matriz +1/-1

$$E = k m n e^{-\lambda S}$$

En cambio, para un HSP con $S = 258$ (86% de identidad sobre 300 residuos), y dados

$k = 0.1$; $n = 300$; $m = 5 \times 10^7$; y $\lambda = 1.58$

$$E = 0.1 \times 300 \times 5e^7 \times e^{-1.58 \times 258} \approx 1.381302e^{-168}$$

Que es un hit altamente significativo, ya que la probabilidad de encontrar un HSP con un score de 258 por azar en la base de datos de 5×10^7 secuencias es de:

$$P = 1 - e^{-E} = 1 - \exp(-1.381302e^{-168}) \approx 0$$

NCBI-BLAST: Basic Local Alignment Search Tool

BLAST consta de una familia de programas, los cuales seleccionaremos en función de:

1. el tipo de secuencia problema (nt | p) [nucleótido | proteína]
2. el tipo de secuencia de la base de datos ainterrogar (nt | p)

Los 5 principales son:

BLASTN (nt-nt), **BLASTP** (p-p), **BLASTX** (translated nt-p),
TBLASTN (p-translated nt), usado en mapeo de prots contra DNA genómico
TBLASTX (translated nt - translated nt) usado en la predicción de genes

y variantes de BLASTP como **PSI-** y **PHI-BLAST**, **DELTA-BLAST** ...

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. [Learn more](#)

NEWS

BLAST+ 2.12.0 is here!

We have made some improvements to how BLAST multi-threads and the amount of memory required by makeblastdb.

Tue, 13 Jul 2021 12:00:00 EST

[More BLAST news...](#)

Web BLAST

<https://blast.ncbi.nlm.nih.gov/Blast.cgi>

Nucleotide BLAST

nucleotide → nucleotide

blastx

translated nucleotide → protein

tblastn

protein → translated nucleotide

Protein BLAST

protein → protein

BLAST: Basic Local Alignment Search Tool

· Anatomía de un reporte de NCBI-BLAST estándar

BLASTP 2.2.13 [Nov-27-2005]

1

Reference:

Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.

RID: 1141782136-12667-92041342765.BLASTPQ4

Database: All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples 3,420,754 sequences; 1,167,289,757 total letters

If you have any problems or questions with the results of this search please refer to the [BLAST FAQ](#)

[taxonnavy_reports](#)

Query: human_myoglobin

Length=154

1.- Encabezado. Indica el programa de BLAST y su versión, con la fecha

Request ID

Indica la BD sobre la que se hizo la búsqueda, junto con el no. de secs contenida en ella y el no. de caracteres

Indica cual fue la query y su longitud

2.- Resumen gráfico de distribución de hits con respecto a la query.

escala de color que indica el score de los HSPs

Las barras indican la distribución de los HSPs (coordenadas) con respecto a la secuencia problema (query), indicando en una escala de color el score de los alns. medidos en bits

© Pablo Vinuesa 2025. @pvinmex; vinuesa[at]ccg[dot]unam[dot]mx; <http://www.ccg.unam.mx/~vinuesa/>

Licencia Creative Commons 4.0, no comercial Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) <https://creativecommons.org/licenses/by-nc/4.0/>

BLAST: Basic Local Alignment Search Tool

• Anatomía de un reporte de NCBI-BLAST estándar

3. **Resúmenes de 1 línea.** Indican el nombre de la sec. junto con el score más alto y E value más bajo encontrado para un HSP o grupo de HSPs

Related Structures

Sequences producing significant alignments:	Score (Bits)	E Value	
gi 4885477 ref NP_005359.1 myoglobin [Homo sapiens] > gi 4495...	316	6e-86	Gene Info
gi 62511907 gb AA084516.1 myoglobin transcript variant 1 [Homo	315	1e-85	
gi 386872 gb AA05995.1 myoglobin	315	1e-85	
gi 229361 prf I711658B myoglobin	313	4e-85	
gi 127683 sp P02145 MYG_PANTR Myoglobin	312	9e-85	
gi 51317414 sp P62735 MYG_HYLEY Myoglobin > gi 51317413 sp P62734	311	1e-84	Structures
gi 127656 sp P02147 MYG_GORBB Myoglobin	311	2e-84	
gi 229360 prf I711658A myoglobin	311	2e-84	
gi 55728442 emb CAH0096.1 hypothetical protein [Pongo pygmaeus	310	5e-84	
gi 230638 pdb 2MM1 Myoglobin Mutant With Lys 45 Replaced By...	309	6e-84	
gi 127689 sp P02148 MYG_PONFY Myoglobin > gi 229570 prf I761377A	308	2e-83	
gi 62901707 sp P68086 MYG_BRVPA Myoglobin > gi 62901706 sp P68...	300	4e-81	

BLAST: Basic Local Alignment Search Tool

• Anatomía de un reporte de NCBI-BLAST estándar


4. **Alineamientos.** Representan la parte más voluminosa del reporte. Además de la información estadística, indica las coordenadas de inicio y fin de las secuencias query y subject. Si la búsqueda involucra secuencias de DNA, también se indica direccionalidad de las hebras Q/S (plus/plus; plus/minus).

>gi|47523546|ref|NP_999401.1| myoglobin [Sus scrofa]
gi|127688|sp|P02189|MYG_PTG Myoglobin
gi|164547|gb|AA031073.1| myoglobin

Length=154
normalized score
raw score

Score = 296 bits (758), Expect = 5e-80
Identities = 144/154 (93%), Positives = 148/154 (96%), Gaps = 0/154 (0%)

Query	1	MGLSDGEWQLVLNVWGKVEADI PGHGQEVLRIRLFGHPETLEKFDKFKHLKSEDEMKASE	60
Sbjct	1	MGLSDGEWQLVLNVWGKVEAD+ GHGQEVLRIRLFGHPETLEKFDKFKHLKSEDEMKASE	60
Query	61	DLKKHGATVLTALGGILKKKGHHEARIKPLAQSHATKKHIVPKYLEFISECIIQVLQSKH	120
Sbjct	61	DLKKHG TVLTALGGILKKKGHHEAB+ PLAQSHATKKHIVPKYLEFISE IIQVLQSKH	120
Query	121	PGDFGADAQGAMNKALELFRKDMASNYKELGFQG 154	
Sbjct	121	PGDFGADAQGAM+KALELFR DMA+ YKELGFQG 154	



BLAST+ desde la línea de comandos

Talleres Internacionales de Bioinformática

TIB2025

Pablo Vinuesa (vinuesa@ccg.unam.mx; @pvinmex)

Centro de Ciencias Genómicas UNAM

<http://www.ccg.unam.mx/~vinuesa/>

Mini-tutorial de uso de [BLAST y] BLAST+ desde la línea de comandos:

1. Generación de bases de datos (indexadas) mediante [formatdb y] makeblastdb

2. Interrogación de bases de datos mediante [blastall -p [blastn|blastp|blastx|tblastn|tblastx] y] blastn, blastp, blastx, delta-blast ...

3. Recuperación de secuencias de una base de datos usando Id's y [fastacmd o] blastdbcmd

Documentación de BLAST+ en NCBI

<http://www.ncbi.nlm.nih.gov/books/NBK1762/>

<http://www.ncbi.nlm.nih.gov/books/NBK52640/>

<http://www.ncbi.nlm.nih.gov/books/NBK279690/>

© Pablo Vinuesa 2025. @pvinmex; vinuesa[at]ccg[dot]unam[dot]mx;
<http://www.ccg.unam.mx/~vinuesa/>

Licencia Creative Commons 4.0, no comercial
Attribution-NonCommercial 4.0 International (CC BY-NC 4.0)
<https://creativecommons.org/licenses/by-nc/4.0/>

FORMATEO DE ARCHIVOS FASTA PARA
GENERAR BASES DE DATOS INTERROGABLES CON BLAST+

- BLAST usa **bases de datos indexadas** para acelerar la operación de búsqueda.
- Existen diversas bases de datos pre-compiladas y formateadas. La más general y extensa es la “nr” o no-redundante. Hay muchas más como: est, wgs, pat, pdb, microbial genomes o env_nt.
- También es posible generar bases de datos propias usando el programa [**formatdb** o] **makeblastdb**. Descárgalo desde <ftp://ftp.ncbi.nih.gov/blast/> junto con los demás binarios de la suite de programas BLAST+. [en ubuntu: apt-get install ncbi-blast+ (blast2 es legacy-blast)]
- Para generar una base de datos se utilizan secuencias en formato FASTA, y con una **sintaxis de identificador NCBI canónica**. Por ejemplo:

lc|integer

lc|string

gnl|yourDB|ID

}

estos son los formatos de las cabeceras FASTA para generar bases de datos de secuencias localmente.

Puedes ver más ejemplos aquí:

http://ncbi.github.io/cxx-toolkit/pages/ch_demo#ch_demo.id1_fetch.html_ref_fasta

Este **identificador es esencial para un correcto indexado de la BD** y poder recuperar secuencias de la BD usando listas de identificadores.

Introducción a BLAST+ desde la línea de comandos

REFERENCIAS CLAVES:

1: Boratyn GM, Camacho C, Cooper PS, Coulouris G, Fong A, Ma N, Madden TL, Matten WT, McGinnis SD, Merezhuk Y, RaytseIis Y, Sayers EW, Tao T, Ye J, Zaretskaya I.BLAST: a more efficient report with usability improvements. Nucleic Acids Res.2013 Jul;41(Web Server issue):W29-33. doi: 10.1093/nar/gkt282. Epub 2013 Apr 22. PubMed PMID: 23609542; PubMed Central PMCID: PMC3692093.

2: Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL.BLAST+: architecture and applications. BMC Bioinformatics. 2009 Dec 15;10:421. doi: 10.1186/1471-2105-10-421. PubMed PMID: 20003500; PubMed Central PMCID: PMC2803857.

Para correr un programa de la suite BLAST+ necesitamos esencialmente 2 cosas:

- Una base de datos a interrogar, adecuadamente formateada
- Una o más secuencias problema con las que buscaremos homólogos en la base de datos llamando a los programas adecuados en función del tipo de secuencia problema (DNA o proteína) y de la base de datos.

Ejemplos de uso de programas de la suite de programas BLAST+

1) formateo de la base de datos

makeblastdb -in sequences4blastdb.fna **-dbtype** nucl **-parse_seqids**

2) ejecutamos una búsqueda con blastn sobre la base de datos recién formateada

blastn -query query_seqs.fas **-db** sequences4blastdb.fna **-out** 16S_out.tab **-outfmt** 6 **-max_target_seqs** 1

3) recuperamos los hits usando blastdbcmd

blastdbcmd -db sequences4blastdb.fna **-entry** my_hits.list

Introducción a BLAST+ desde la línea de comandos

Ayuda desde la línea de comandos:

1. Ayuda en formato condensado:

Programa -h (por ejemplo: blastn -h)

2. Ayuda detallada

Programa -help (por ejemplo: blastp -help)

Conviene revisar además el **BLAST Command Line Applications User Manual** en: <http://www.ncbi.nlm.nih.gov/books/NBK1763/>

Sigue un resumen de algunos comandos básicos y sus opciones, comparando el “blast viejo o legacy blast” con blast+ (actual)

BLAST	BLAST+	Descripción
formatdb	makeblastdb	
-i	-in	Archivo de entrada con secuencias
-p T/F	-dbtype prot/nucl	Mol type
-o T	-parse_seqids	Parsea e indexa seq IDs
-n	-out	Nombre de base para archivos de salida

BLAST+ - el nuevo BLAST escrito en C++

Continuación (ver blast[npq...] -h para despliegue de opciones

BLAST	BLAST+	Descripción
blastall	blastn, blastp,...	
-p	No existe	blastn, blastp, blastx, tblastn, ...
-i	-query	Archivo de entrada
-d	-db	Base de datos de blast
-o	-out	Nobre de archivos de salida
-m	-outfmt	Formato salida; TAB: 6 == m 8
-e	-evaluate	Punto de corte para valor de Expectancia
-v	-num_descriptions	Máximo número de descripciones - hits
-b	-num_alignments	Número máximo de alineamientos
-a	-num_threads	No. de cores a usar
	-max_target_seqs	No. max. de secuencias y descripciones
-F F	-dust no -seg no	Deshabilitar filtrado de regiones de baja complejidad; DNA:dust AA:seg

BLAST+ - el nuevo BLAST escrito en C++

BLAST	BLAST+	Descripción
fastacmd	blastdbcmd	
-d	-db	Base de datos de blast
-s	-entry	Cadena de búsqueda
-DB 1	-entry all	DB dump en formato FASTA

Ejemplos de uso de programas de la suite de programas BLAST+

- # 1) Formateo de la base de datos
makeblastdb -in sequences4blastdb.fna -dbtype nucl -parse_seqids
- # 2) ejecutamos una búsqueda con blastn sobre la base de datos recién formateada
blastn -query query_seqs.fas -db sequences4blastdb.fna -out 16S_out.tab -outfmt 6 \ -max_target_seqs 1
- # 3) recuperamos los hits usando blastdbcmd
blastdbcmd -db sequences4blastdb.fna -entry my_hits.list

Ya es hora de hacer unos ejercicios con datos reales...

Ejercicios: formateo de bases de datos de nt y aa con makeblastdb y búsquedas locales con blastn, blastp y blastx

- I. Formateo de base de datos de secuencias 16S de *Mycobacterium* spp. y búsqueda en ella de homólogos mediante blastn
 - 1) Descargar el archivo 16S_4blastN.tgz de la página del curso
 - 2) Descomprimirlo y abrir el tarro con: tar -xvzf 16S_4blastN.tgz
 - 3) Construiremos la base de datos con las secuencias disponibles en el archivo 16S_seqs4_blastDB.fna. Primero que nada averigüen:
 - 3.1 ¿ cuantas secuencias tiene; cuantas especies representa?
 - 3.2 ¿qué información contienen los identitificadores (el fasta header) ?
 - 3.3 ¿ es su formato adecuado para un indexado correcto?Usa la línea de comandos para dar respuesta a estar preguntas
 - 1) ¿Qué línea de comando usarías para un generar una base de datos con el archivo 16S_seqs4_blastDB.fna para que esté indexado?
 - 1) ¿Cómo clasificarías las secuencias contenidas en el archivo 16S_problema.fna ?
 - 2) Recupera los 10 hits más próximos a cada secuencia problema de la base de datos para su posterior alineamiento; filtra aquellos hits con >= 98.5% de identidad

Ejercicios: continuación

- II. Formateo de base de datos de secuencias de integrones bacterianos y descubrimiento y anotación de genes (cassettes) amplificados de cepas de *E. coli* recuperadas por Jazmín Madrigal del río Apatlaco, Mor. México.
 - 1) Descargar el archivo gene_discovery_and_annotation_using_blastx.tgz de la página
 - 2) Descomprimirlo y abrir el tarro con:
tar -xvzf gene_discovery_and_annotation_using_blastx.tgz
 - 3) Construiremos la base de datos con las secuencias disponibles en el archivo integron_cassettes4blastdb.faa. Primero que nada averigüen:
 - 3.1 ¿ cuántas secuencias tiene; cuantas especies representa?
 - 3.2 ¿qué información contienen los identitificadores (el fasta header) ?
 - 3.3 ¿ es su formato adecuado para un indexado correcto?Usa la línea de comandos (shell) para dar respuesta a estas preguntas
 - 4) ¿Qué comando usarías para un generar una base de datos con el archivo *4blastdb.faa para que esté indexado?
 - 5) ¿Qué comandos usarías para identificar y anotar los genes que pudieran estar codificados en las secuencias contenidas en el archivo 3cass_amplicons.fna?
 - 6) Recupera los 10 hits más próximos a cada secuencia problema de la base de datos para su posterior alineamiento.

Campos del formato tabular -outfmt 6 de BLAST+ (-m 8 legacy)

- Como ya vimos, la opción -outfmt 6 de blast[n|p|x] especifica una salida en formato tabular, con los campos separados por tabuladores.
- Estos datos (líneas) se pueden parsear fácilmente usando AWK o comandos de UNIX como:

```
# imprime sólo hits con %ID > 95% y aln_len > 500
awk '$3 > 95.0 && $4 > 500 { print "$1\t$2" }' blast_fmt6.out
# obtén una lista no redundante de hits
cut -f2 blast_output.txt | sort -u
```
- Los 12 campos o columnas estándar son las siguientes: (-outfmt 7 los imprime)
 - 1: query name
 - 2: subject name
 - 3: percent identities
 - 4: alignment length
 - 5: number of mismatched positions
 - 6: number of gap positions
 - 7: query sequence start
 - 8: query sequence end
 - 9: subject sequence start
 - 10: subject sequence end
 - 11: e-value
 - 12: bit score